Seminario

Testing gene-environment interaction in generalized linear mixed models with family data

20 de noviembre de 2017

Sección 1

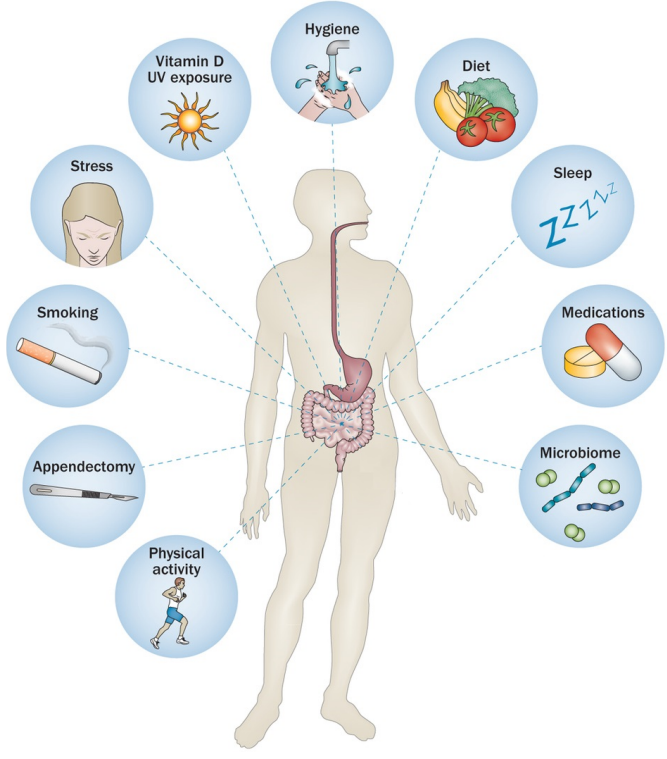
Introduction

Gene, Environment and Disease

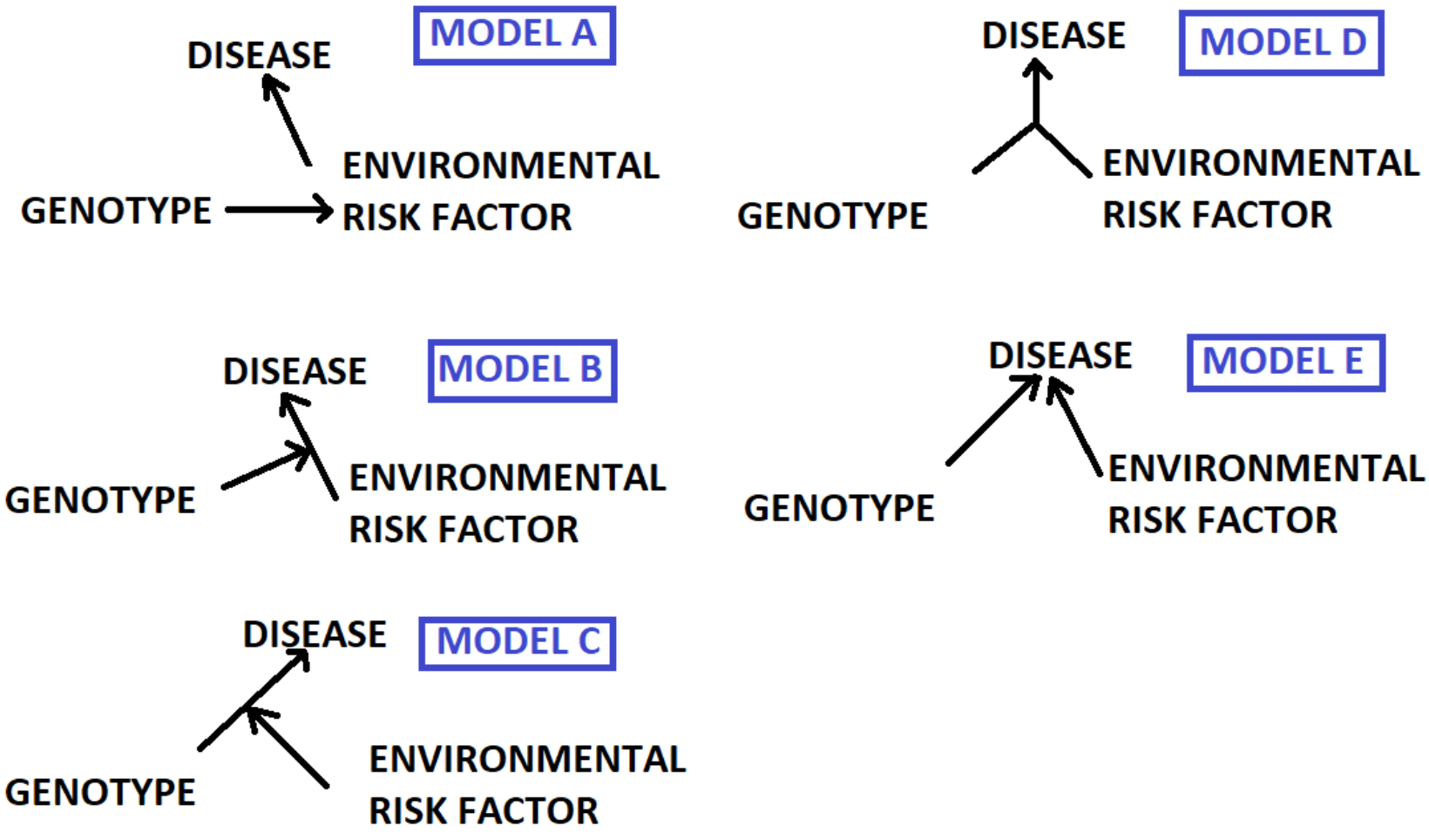
DISEASE (PHENOTYPE)

GENES

ENVIRONMENT

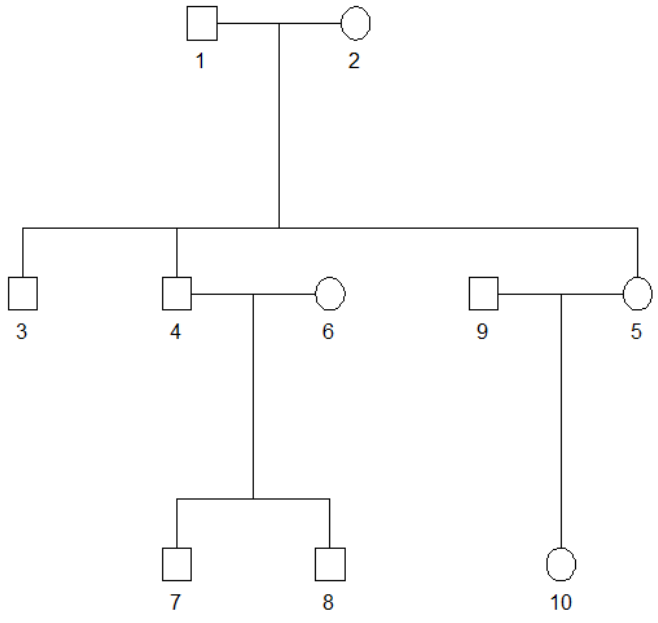


Gene-environment interaction



Family data and kinship matrix

PEDIGREE



KINSHIP MATRIX

Subject	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.50	0.50	0.50	0.0	0.250	0.250	0.0	0.250
2	0.00	1.00	0.50	0.50	0.50	0.0	0.250	0.250	0.0	0.250
3	0.50	0.50	1.00	0.50	0.50	0.0	0.250	0.250	0.0	0.250
4	0.50	0.50	0.50	1.00	0.50	0.0	0.500	0.500	0.0	0.250
5	0.50	0.50	0.50	0.50	1.00	0.0	0.250	0.250	0.0	0.500
6	0.00	0.00	0.00	0.00	0.00	1.0	0.500	0.500	0.0	0.000
7	0.25	0.25	0.25	0.50	0.25	0.5	1.000	0.500	0.0	0.125
8	0.25	0.25	0.25	0.50	0.25	0.5	0.500	1.000	0.0	0.125
9	0.00	0.00	0.00	0.00	0.00	0.0	0.000	0.000	1.0	0.500
10	0.25	0.25	0.25	0.25	0.50	0.0	0.125	0.125	0.5	1.000

Notation

- Consider a sample of N independent families.
- n_i: number of members in the ith family ($i = 1, \ldots, N$).
- Y_{ij}: reponse variable for the phenotype (discrete or continuous).
- $\boldsymbol{X}_{ij} = (X_{ij}^1, \ldots, X_{ij}^p)^T$: p non-environmental covariates.
- $\boldsymbol{G}_{ij} = (G_{ij}^1, \ldots, G_{ij}^q)^T$: q observed genotypes at certain targeted genetic marker loci.
- E_{ij}: some environmental exposure factor of interest.
- $\boldsymbol{S}_{ij} = (E_{ij}G_{ij}^1, \ldots, E_{ij}G_{ij}^q)^T$: GE interaction.

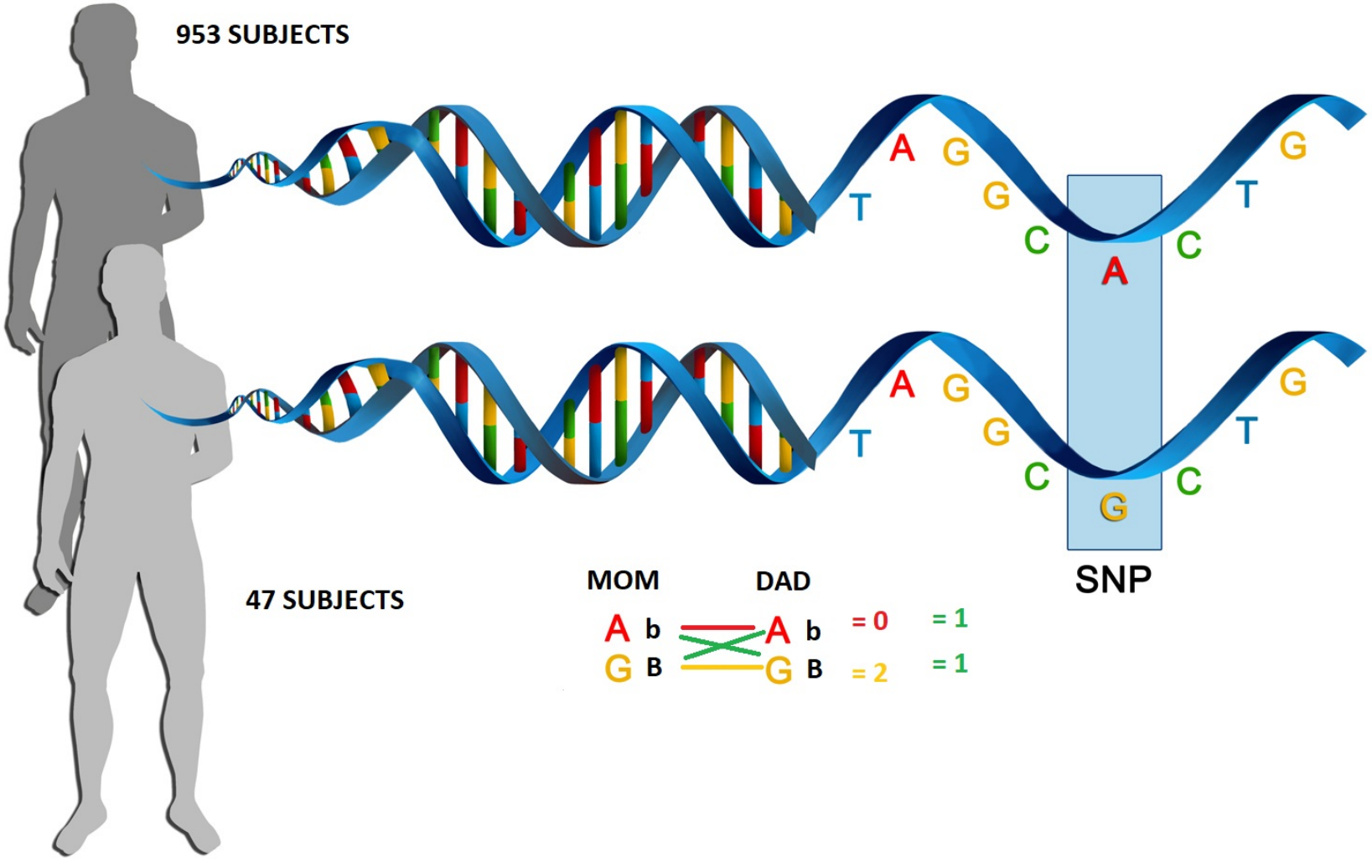
Observed genotypes - $\boldsymbol{G}_{ij}$ -

Each column in $\boldsymbol{G}_{ij}$ is a Single Nucleotide Polymorphism (SNP). The genetic information is represented according to the following codification:

$$G_{ij}^{k} = \begin{cases} 2, & \text{subject } j \text{ at } i\text{th family is homozygous BB} \\ 1, & \text{subject } j \text{ at } i\text{th family is heterozygous Bb or bB} \\ 0, & \text{subject } j \text{ at } i\text{th family is homozygous bb,} \end{cases}$$

with $k = 1, \ldots, q$. B and b represent the dominant and recessive alleles, respectively. In addition, B is the allele that occurs at minor frequency.

Observed genotypes - G_{ij} -



Sección 2

Model

Generalized linear mixed model (GLMM) for Subject j at ith family

$$
\begin{aligned}
g\left[E(Y_{ij}|\alpha_{ij})\right] &= \boldsymbol{X}_{ij}^T\beta_1 + E_{ij}\beta_2 + \boldsymbol{G}_{ij}^T\boldsymbol{\theta} + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma} + \alpha_{ij}, \quad (1)\\
Var(Y_{ij}|\alpha_{ij}) &= \phi\omega_{ij}^{-1}\nu\left[E(Y_{ij}|\alpha_{ij})\right],\\
\boldsymbol{\alpha}_i &= (\alpha_{i1},\ldots,\alpha_{in_i})^T \sim N(\mathbf{0}, 2\sigma^2\boldsymbol{\Phi}_i),
\end{aligned}
$$

where,

- $g(.)$: monotone known function.
- $Y_{ij}|\alpha_{ij}$ follows a distribution in the exponential family.
- $\nu(.)$: known function.
- ϕ: a scale parameter that may be known or may need to be estimated.
- ω_{ij}: known weights (commonly equal to 1).
- $\boldsymbol{\Phi}_i$: the kinship matrix and σ^2 is a parameter to be estimated.

Examples of link functions

Family	Link	$g(\cdot)$	Trait	$\nu(\cdot)$
Binomial	Logit	$\ln\left(\frac{\mu}{1-\mu}\right)$	Binary	$\mu(1-\mu)$
Gaussian	Identity	μ	Continuous	ϕ
Gamma	Inverse	$1/\mu$	Continuous	$\phi\mu^2$
Inverse.gaussian	Inverse squared	$1/\mu^2$	Continuous	$\phi\mu^3$
Poisson	Log	$\ln(\mu)$	Count	μ
Quasi	Identity	μ	Continuous	ϕ
Quasibinomial	Logit	$\ln\left(\frac{\mu}{1-\mu}\right)$	Binary	$\phi\mu(1-\mu)$
Quasipoisson	Log	$\ln(\mu)$	Count	$\phi\mu$

GLMM for ith Family

$$g\left(\boldsymbol{\mu}_i^b\right) = \boldsymbol{X}_i\beta_1 + \boldsymbol{E}_i\beta_2 + \boldsymbol{G}_i\boldsymbol{\theta} + \boldsymbol{S}_i\boldsymbol{\gamma} + \boldsymbol{K}_i\boldsymbol{b}_i, \qquad (2)$$

where,

- $2\boldsymbol{\Phi}_i = \boldsymbol{K}_i\boldsymbol{K}_i^T$ (Cholesky decomposition).
- $\boldsymbol{\alpha}_i = \boldsymbol{K}_i\boldsymbol{b}_i$, with $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$ and $\boldsymbol{I}_{n_i}$: identity matrix.

- $\boldsymbol{\mu}_i^b = E(\boldsymbol{Y}_i|\boldsymbol{b}_i)$.
- $g(\boldsymbol{\mu}_i^b) = (g(\mu_{i1}^b), \ldots, g(\mu_{in_i}^b))^T$.
- $\boldsymbol{X}_i = [\boldsymbol{X}_{i1} \ldots \boldsymbol{X}_{in_i}]^T$.
- $\boldsymbol{G}_i = [\boldsymbol{G}_{i1} \ldots \boldsymbol{G}_{in_i}]^T$.
- $\boldsymbol{E}_i = (E_{i1}, \ldots, E_{in_i})^T$.
- $\boldsymbol{S}_i = [\boldsymbol{S}_{i1} \ldots \boldsymbol{S}_{in_i}]^T$.

General GLMM

$$g\left(\mu^b\right) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{S}\gamma + \boldsymbol{K}\boldsymbol{b}, \quad (3)$$

where

- $\mu^b = E(\boldsymbol{Y}|\boldsymbol{b})$.
- $\boldsymbol{X} = [\boldsymbol{X}_1 \dots \boldsymbol{X}_N]^T$.
- $\boldsymbol{G} = [\boldsymbol{G}_1 \dots \boldsymbol{G}_N]^T$.
- $\boldsymbol{E} = (E_1, \dots, E_N)^T$
- $\boldsymbol{S} = [\boldsymbol{S}_1 \dots \boldsymbol{S}_N]^T$.
- $\boldsymbol{K} = \text{diag}\{\boldsymbol{K}_1 \dots \boldsymbol{K}_N\}$.
- $\boldsymbol{b} = [\boldsymbol{b}_1 \dots \boldsymbol{b}_N]^T$.
- $\tilde{\boldsymbol{X}} = [\boldsymbol{X}\ \boldsymbol{E}]^T$.
- $\beta = (\beta_1^T, \beta_2)^T$.

We are interested in testing the hypthotesis $H_0 : \gamma = \mathbf{0}$.

Important facts

- If γ is treated as a fixed vector and the null hyphotesis is tested with a q degrees of freedom score test, it can result in loss of power (Lin *et. al.*, 2013).

Important facts

- If γ is treated as a fixed vector and the null hyphotesis is tested with a q degrees of freedom score test, it can result in loss of power (Lin *et. al.*, 2013).
- Another common strategy is to use a single SNP at time to test GE interaction.

Important facts

- If γ is treated as a fixed vector and the null hyphotesis is tested with a q degrees of freedom score test, it can result in loss of power (Lin *et. al.*, 2013).
- Another common strategy is to use a single SNP at time to test GE interaction.
- Assume γ as a random vector following a multivariate normal distribution $N(\mathbf{0}, \tau \boldsymbol{I}_q)$ and to test the equivalent null hypothesis $H_0 : \tau = 0$.

Important facts

- If γ is treated as a fixed vector and the null hyphotesis is tested with a q degrees of freedom score test, it can result in loss of power (Lin *et. al.*, 2013).
- Another common strategy is to use a single SNP at time to test GE interaction.
- Assume γ as a random vector following a multivariate normal distribution $N(\mathbf{0}, \tau \boldsymbol{I}_q)$ and to test the equivalent null hypothesis $H_0 : \tau = 0$.

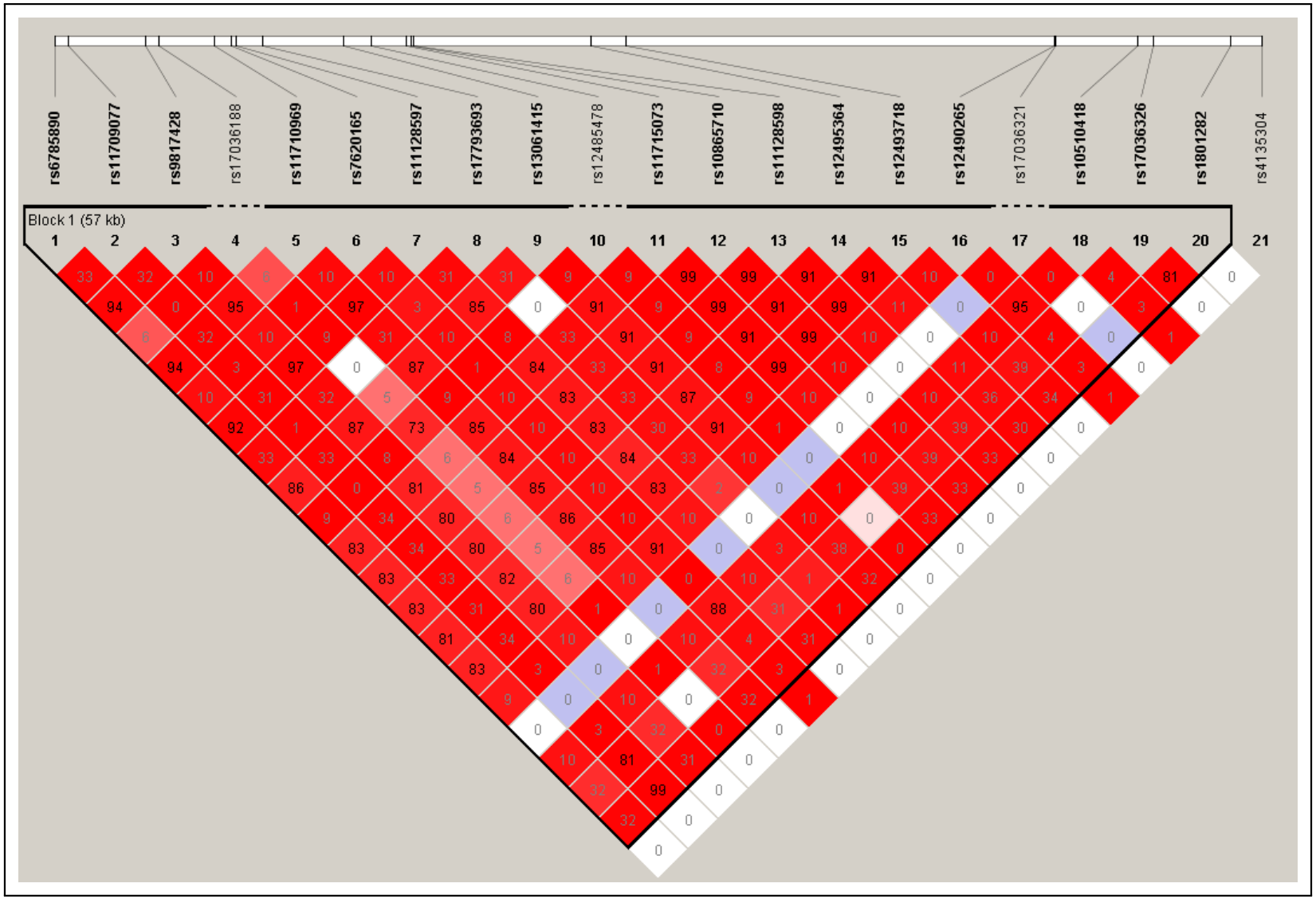
$$g\left(\boldsymbol{\mu}^{b,\gamma}\right) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{G}\boldsymbol{\theta} + \underbrace{\boldsymbol{S}\gamma + \boldsymbol{K}\boldsymbol{b}}_{random}$$

Important facts

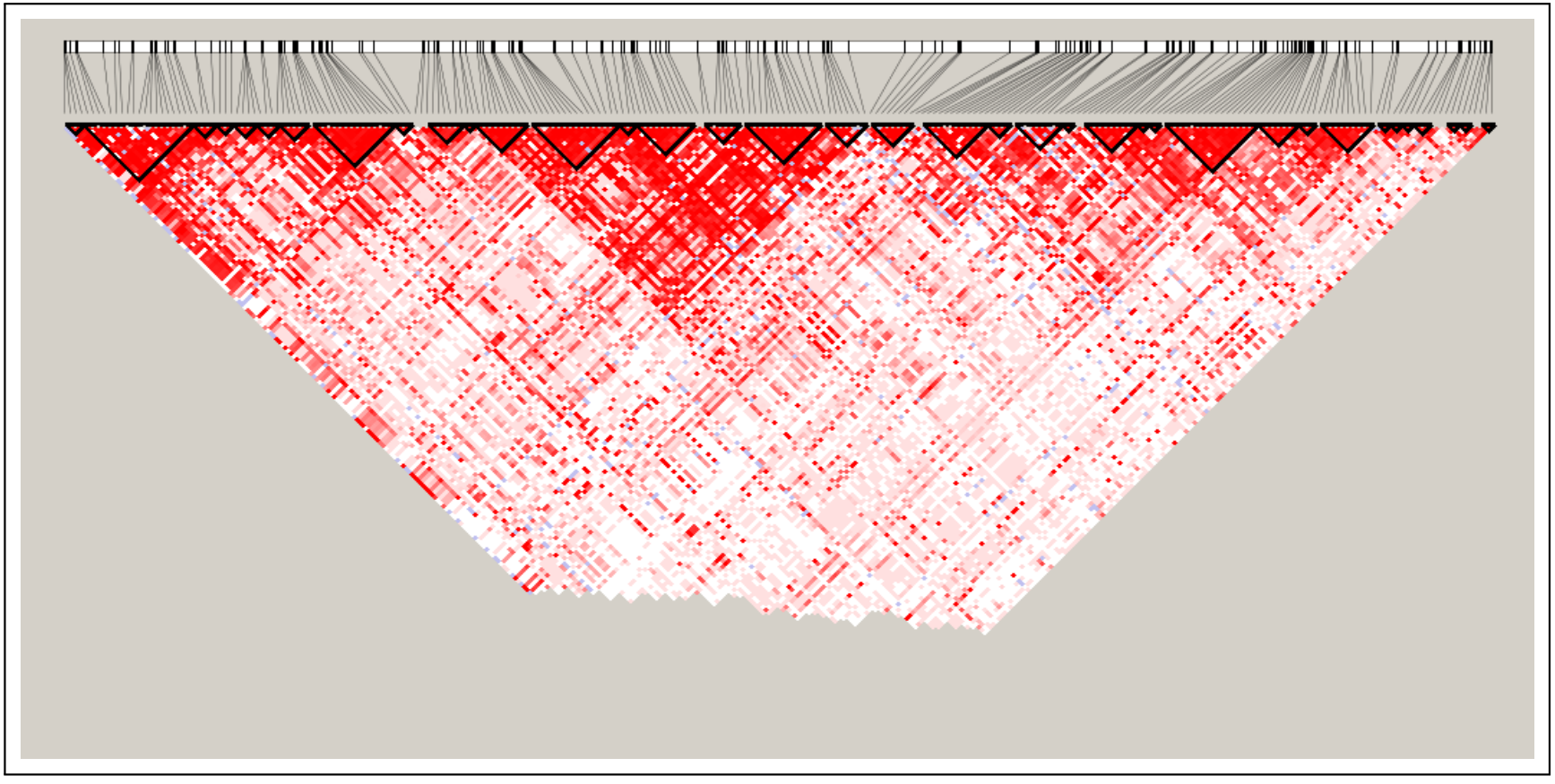
- If γ is treated as a fixed vector and the null hyphotesis is tested with a q degrees of freedom score test, it can result in loss of power (Lin *et. al.*, 2013).
- Another common strategy is to use a single SNP at time to test GE interaction.
- Assume γ as a random vector following a multivariate normal distribution $N(\mathbf{0}, \tau \boldsymbol{I}_q)$ and to test the equivalent null hypothesis $H_0 : \tau = 0$.

$$g\left(\boldsymbol{\mu}^b\right) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{S}\gamma + \underbrace{\boldsymbol{K}\boldsymbol{b}}_{random}$$

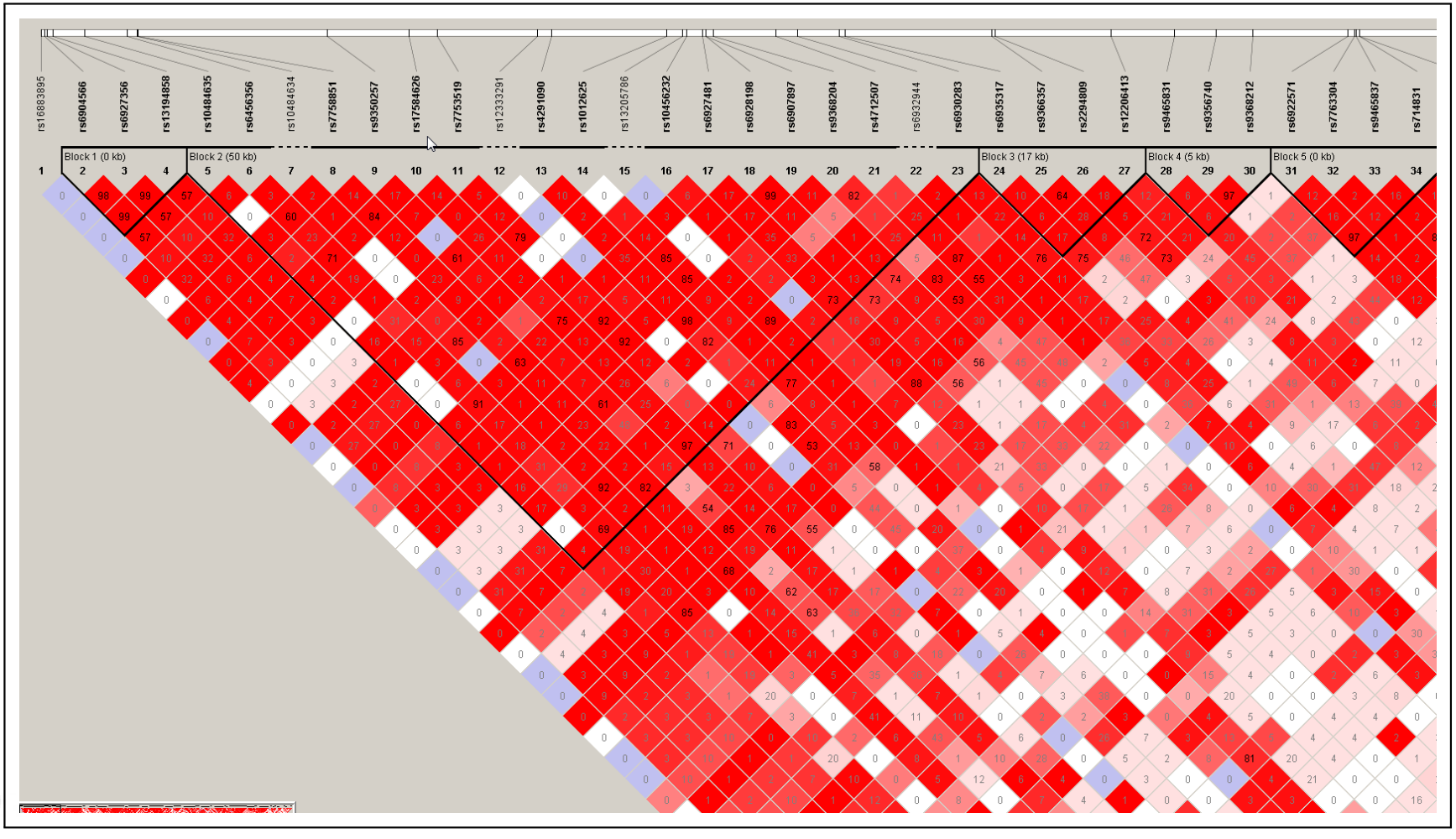
Linkage disequilibrium (LD) -PPARG-



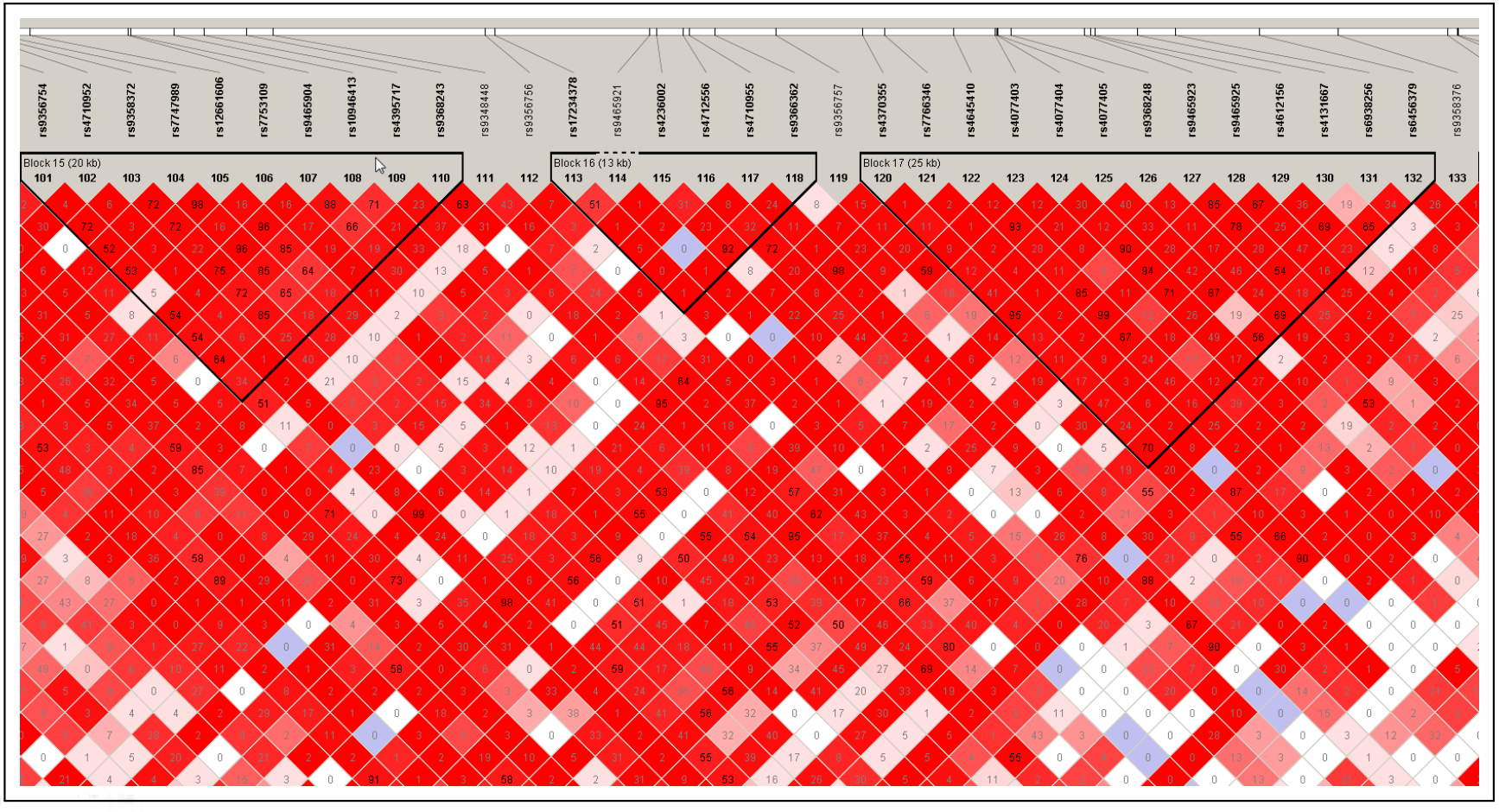
Linkage disequilibrium (LD) -CDKAL1-



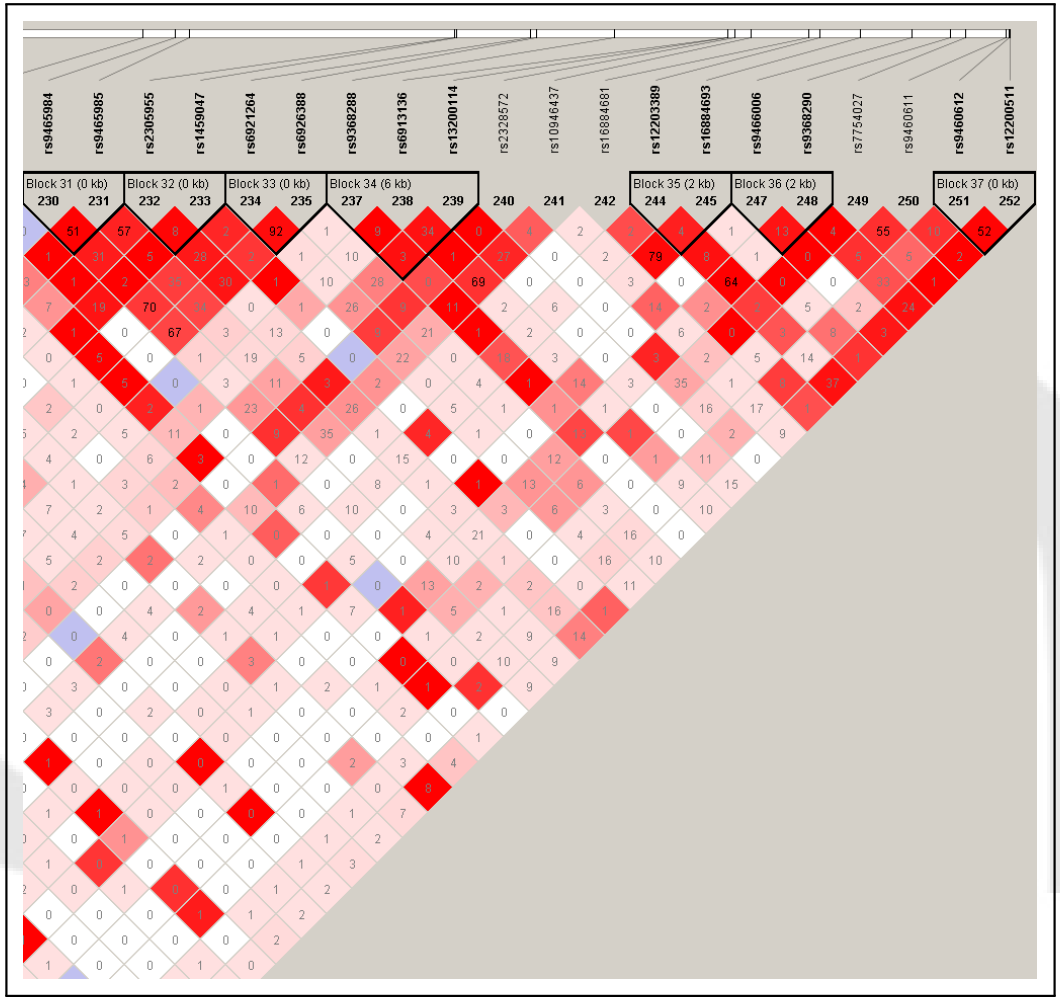
Linkage disequilibrium (LD) -CDKAL1-



Linkage disequilibrium (LD) -CDKAL1-



Linkage disequilibrium (LD) -CDKAL1-



Remedial considerations to face LD

- Given the high correlation among the SNPs in $\boldsymbol{G}$, Lin *et. al.* (2013) proposed (for independet subjects) to penalize the estimation of parameter $\boldsymbol{\theta}$ by using ridge regression and introducing a penalization term λ in the quasi-likelihood function. The tunning parameter λ is selected using generalized cross validation (Fu, 2005).

Remedial considerations to face LD

- Given the high correlation among the SNPs in $\boldsymbol{G}$, Lin *et. al.* (2013) proposed (for independet subjects) to penalize the estimation of parameter $\boldsymbol{\theta}$ by using ridge regression and introducing a penalization term λ in the quasi-likelihood function. The tunning parameter λ is selected using generalized cross validation (Fu, 2005).
- Shen *et. al.* (2013) proposed for generalized linear models (GLM) that ridge regression is equivalent to assume the penalized parameters as independent random variables.

Remedial considerations to face LD

It is also equivalent for GLMM and assuming $\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2_\theta \boldsymbol{I}_q)$, it is possible to show that $\lambda = \phi/\sigma^2_\theta$.

Remedial considerations to face LD

It is also equivalent for GLMM and assuming $\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2_\theta \boldsymbol{I}_q)$, it is possible to show that $\lambda = \phi/\sigma^2_\theta$.

$$g\left(\boldsymbol{\mu}^{b,\gamma,\theta}\right) = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \underbrace{\boldsymbol{G\theta} + \boldsymbol{S\gamma} + \boldsymbol{Kb}}_{random}$$

Remedial considerations to face LD

It is also equivalent for GLMM and assuming $\boldsymbol{\theta} \sim N(\mathbf{0}, \sigma^2_\theta \boldsymbol{I}_q)$, it is possible to show that $\lambda = \phi/\sigma^2_\theta$.

$$g\left(\boldsymbol{\mu}^{b,\theta}\right) = \tilde{\boldsymbol{X}}\beta + \underbrace{\boldsymbol{G\theta}}_{random} + S\gamma + \underbrace{\boldsymbol{Kb}}_{random}$$

Sección 3

Null model estimation

Null model estimation

$$g\left(\boldsymbol{\mu}^{\boldsymbol{d}_1,\boldsymbol{d}_2}\right) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{d}_1 + \boldsymbol{d}_2$$

with $\boldsymbol{d}_1 = \boldsymbol{G}\boldsymbol{\theta}$, $\boldsymbol{d}_2 = \boldsymbol{K}\boldsymbol{b}$.

Null model estimation

$$g\left(\boldsymbol{\mu}^{\boldsymbol{d}_1,\boldsymbol{d}_2}\right) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{d}_1 + \boldsymbol{d}_2$$

with $\boldsymbol{d}_1 = \boldsymbol{G}\boldsymbol{\theta}$, $\boldsymbol{d}_2 = \boldsymbol{K}\boldsymbol{b}$. Breslow and Clayton (1993) proposed a Fisher scoring solution that may be expressed as the iterative solution to the system

$$\begin{bmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^T\boldsymbol{W} & \tilde{\boldsymbol{X}}^T\boldsymbol{W} \\ \boldsymbol{W}\tilde{\boldsymbol{X}} & \frac{\phi}{\sigma^2_\theta}(\boldsymbol{G}\boldsymbol{G}^T)^{-1} + \boldsymbol{W} & \boldsymbol{W} \\ \boldsymbol{W}\tilde{\boldsymbol{X}} & \boldsymbol{W} & \frac{\phi}{\sigma^2}(\boldsymbol{K}\boldsymbol{K}^T)^{-1} + \boldsymbol{W} \end{bmatrix} \begin{pmatrix} \beta \\ \boldsymbol{d}_1 \\ \boldsymbol{d}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}\tilde{\boldsymbol{Y}} \end{pmatrix}$$

- $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\beta + \boldsymbol{d}_1 + \boldsymbol{d}_2 + \varepsilon$: working vector;
- $\varepsilon \sim N(\boldsymbol{0}, \phi\boldsymbol{W}^{-1})$;
- $\boldsymbol{W} = \text{diag}\left\{\omega_{ij} / \left[\nu(\mu^d_{ij})g'(\mu^d_{ij})^2\right]\right\}$.

Ridge

Null model estimation

$$\widehat{\beta} = (\tilde{\boldsymbol{X}}^T \Sigma^{-1} \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \Sigma^{-1} \tilde{\boldsymbol{Y}},$$

$$\begin{pmatrix} \widehat{\boldsymbol{d}}_1 \\ \widehat{\boldsymbol{d}}_2 \end{pmatrix} = \begin{pmatrix} \sigma_\theta^2 \left(\boldsymbol{G}\boldsymbol{G}^T\right) \Sigma^{-1} (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\widehat{\beta}) \\ \sigma^2 \left(\boldsymbol{K}\boldsymbol{K}^T\right) \Sigma^{-1} (\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\widehat{\beta}) \end{pmatrix},$$

with $\Sigma = \sigma_\theta^2 \boldsymbol{G}\boldsymbol{G}^T + \sigma^2 \boldsymbol{K}\boldsymbol{K}^T + \phi \boldsymbol{W}^{-1}$.

Sección 4

GE interaction test

Testing GE interaction

Following the approach developed by Zhang and Lin (2003), we propose as score statitistic the quadratic form

$$U_\tau = U_\tau\left(\widehat{\beta}, \widehat{\pi}\right) = \frac{1}{2}\left\{\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta\right)^T \Sigma^{-1} \boldsymbol{S}\boldsymbol{S}^T \Sigma^{-1}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta\right)\right\}\Big|_{\widehat{\beta}.\widehat{\pi}}.$$

with $\widehat{\pi}$ is the estimator of $\pi = (\sigma_\theta^2, \sigma^2, \phi)^T$.

Testing GE interaction

Following the approach developed by Zhang and Lin (2003), we propose as score statitistic the quadratic form

$$U_\tau = U_\tau\left(\widehat{\beta}, \widehat{\boldsymbol{\pi}}\right) = \frac{1}{2}\left\{\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta\right)^T \Sigma^{-1} \boldsymbol{S}\boldsymbol{S}^T \Sigma^{-1}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\beta\right)\right\}\Big|_{\widehat{\beta}.\widehat{\boldsymbol{\pi}}}.$$

with $\widehat{\boldsymbol{\pi}}$ is the estimator of $\boldsymbol{\pi} = (\sigma_\theta^2, \sigma^2, \phi)^T$.To correct for bias, we use the restricted maximum likelihood (REML) estimators (Breslow and Clayton, 1993) in the GLMM framework to obtain $\widehat{\beta}$ and $\widehat{\boldsymbol{\pi}}$ under the null hypothesis.

Testing GE interaction

Zhang and Lin (2003) showed that under $H_0 : \tau = 0$, U_τ follows approximately a mixture of one degree of freedom independent chi-square distributions. However, for computational ease, we use the Satterthwaite method (Satterthwaite, 1941) to approximate the distribution of U_τ by a scaled chi-square distribution $\kappa\chi^2_\xi$.

Testing GE interaction

Zhang and Lin (2003) showed that under $H_0 : \tau = 0$, U_τ follows approximately a mixture of one degree of freedom independent chi-square distributions. However, for computational ease, we use the Satterthwaite method (Satterthwaite, 1941) to approximate the distribution of U_τ by a scaled chi-square distribution $\kappa\chi^2_\xi$.
When REML estimates are used to calculate U_τ showed that the mean and variance of U_τ can be approximated, respectively, by

$$\text{tr}\left(\boldsymbol{PSS}^T\right)\Big|_{\widehat{\boldsymbol{\pi}}} \quad \text{and} \quad \mathcal{I}_\tau = \frac{1}{2}\left\{\text{tr}(\boldsymbol{PSS}^T\boldsymbol{PSS}^T) - \boldsymbol{J}^T\boldsymbol{M}^{-1}\boldsymbol{J}\right\}\Big|_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\pi}}}$$

Testing GE interaction

$$\boldsymbol{J} = \left(\begin{array}{cc|c} \text{tr}[\boldsymbol{PSS}^T\boldsymbol{PGG}^T] & \text{tr}[\boldsymbol{PSS}^T\boldsymbol{PKK}^T] & \text{tr}[\boldsymbol{PSS}^T\boldsymbol{PW}^{-1}] \end{array} \right)^T$$

$$\boldsymbol{M} = \left[\begin{array}{cc|c} \text{tr}[\boldsymbol{PGG}^T\boldsymbol{PGG}^T] & \text{tr}[\boldsymbol{PGG}^T\boldsymbol{PKK}^T] & \text{tr}[\boldsymbol{PGG}^T\boldsymbol{PW}^{-1}] \\ \text{tr}[\boldsymbol{PKK}^T\boldsymbol{PG}^T\boldsymbol{G}] & \text{tr}[\boldsymbol{PKK}^T\boldsymbol{PKK}^T] & \text{tr}[\boldsymbol{PKK}^T\boldsymbol{PW}^{-1}] \\ \hline \text{tr}[\boldsymbol{PW}^{-1}\boldsymbol{PG}^T\boldsymbol{G}] & \text{tr}[\boldsymbol{PW}^{-1}\boldsymbol{PKK}^T] & \text{tr}[\boldsymbol{PW}^{-1}\boldsymbol{PW}^{-1}] \end{array} \right],$$

and $\boldsymbol{P} = \Sigma^{-1} - \Sigma^{-1}\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}^T\Sigma^{-1}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\Sigma^{-1}$.

Testing GE interaction

Since the mean and variance of $\kappa\chi^2_\xi$ are given by $\kappa\xi$ and $2\kappa^2\xi$, respectively, we obtain the equations $\text{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T) = \kappa\xi$ and $\mathcal{I}_\tau = 2\kappa^2\xi$, where $\widehat{\boldsymbol{P}}$ denotes the matrix $\boldsymbol{P}$ evaluated in $\widehat{\boldsymbol{\pi}}$. By solving these equations, we demonstrate that

$$\kappa = \mathcal{I}_\tau / [2\,\text{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T)]$$

and

$$\xi = 2\left[\text{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T)\right]^2 / \mathcal{I}_\tau.$$

Testing GE interaction

Since the mean and variance of $\kappa\chi^2_\xi$ are given by $\kappa\xi$ and $2\kappa^2\xi$, respectively, we obtain the equations $\text{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T) = \kappa\xi$ and $\mathcal{I}_\tau = 2\kappa^2\xi$, where $\widehat{\boldsymbol{P}}$ denotes the matrix $\boldsymbol{P}$ evaluated in $\widehat{\boldsymbol{\pi}}$. By solving these equations, we demonstrate that

$$\kappa = \mathcal{I}_\tau / [2\,\text{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T)]$$

and

$$\xi = 2\left[\text{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T)\right]^2 / \mathcal{I}_\tau.$$

$$\boxed{\frac{U_\tau}{\kappa} \sim \chi^2_\xi}$$

Sección 5

Simulations

Simulated model

$$E_{ij} = 2 + 0{,}01 Age_{ij} + 0{,}1 I(Female_{ij}) + \gamma_i + \varepsilon_{ij};$$

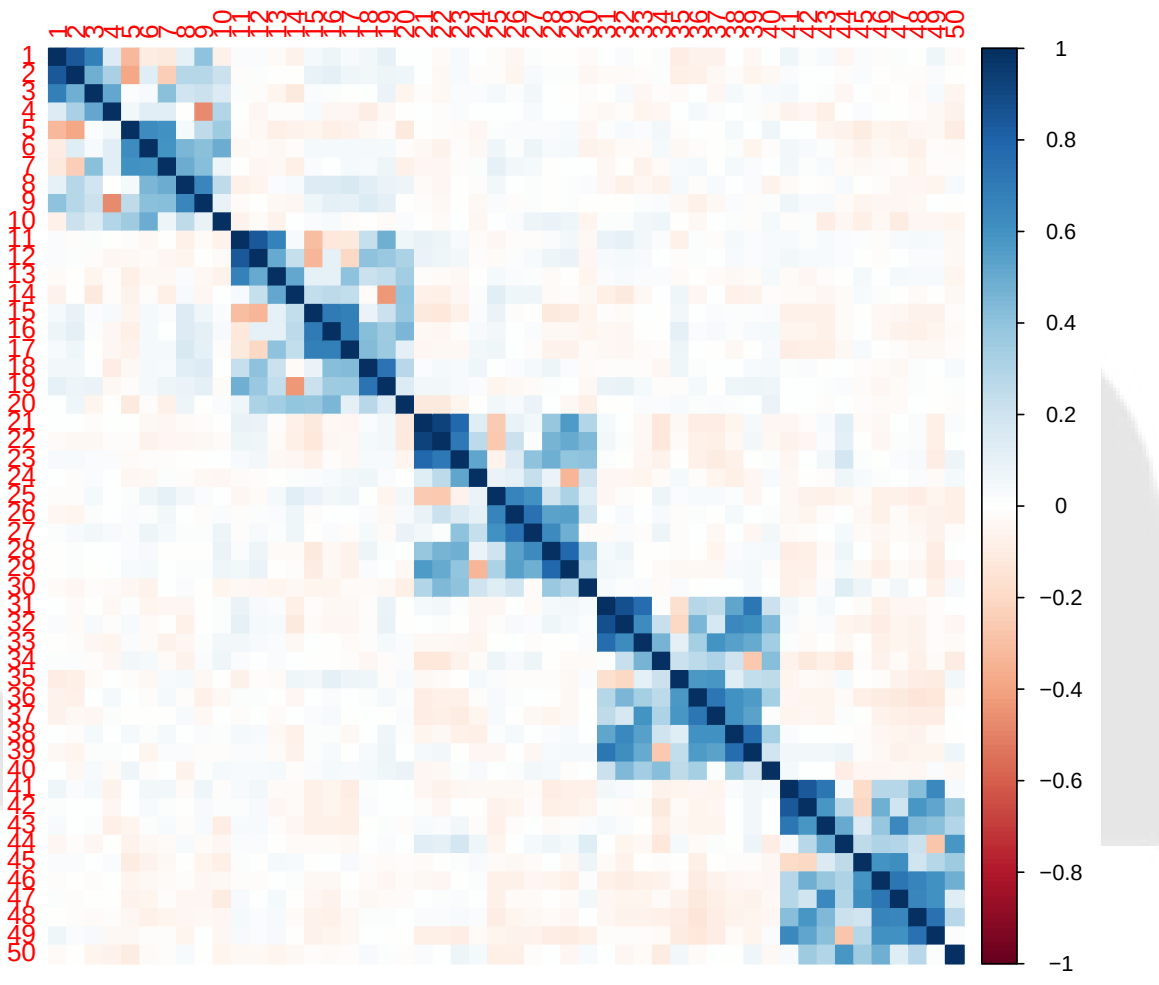
$$\begin{aligned} & \text{logit}\left[P\left(Y_{ij} = 1 | Age_{ij}, Female_{ij}, E_{ij}, G1_{ij}, G2_{ij}\right)\right] \\ = \; & 0{,}1 + 0{,}01 Age_{ij} + 0{,}1 I(Female_{ij}) + 0{,}1 E_{ij} + 0{,}3 G1_{ij} \\ & + 0{,}3 G2_{ij} + \gamma_1 (G1_{ij} \times E_{ij}) + \gamma_2 (G2_{ij} \times E_{ij}) + \alpha_{ij} \end{aligned}$$

with

- $I(\cdot)$ is the indicator function;
- $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i10})^T \sim N(\mathbf{0}, 4\boldsymbol{I}_{10})$;
- $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{i10})^T \sim N(\mathbf{0}, 2\boldsymbol{\Phi}_i)$;
- $\gamma_i \sim N(0, 4)$

- $\boldsymbol{\Phi}_i$ is the kinship matrix corresponding to the aforementioned family pedigree.

Correlation of the 50 simulated SNPs in LD



Type I error

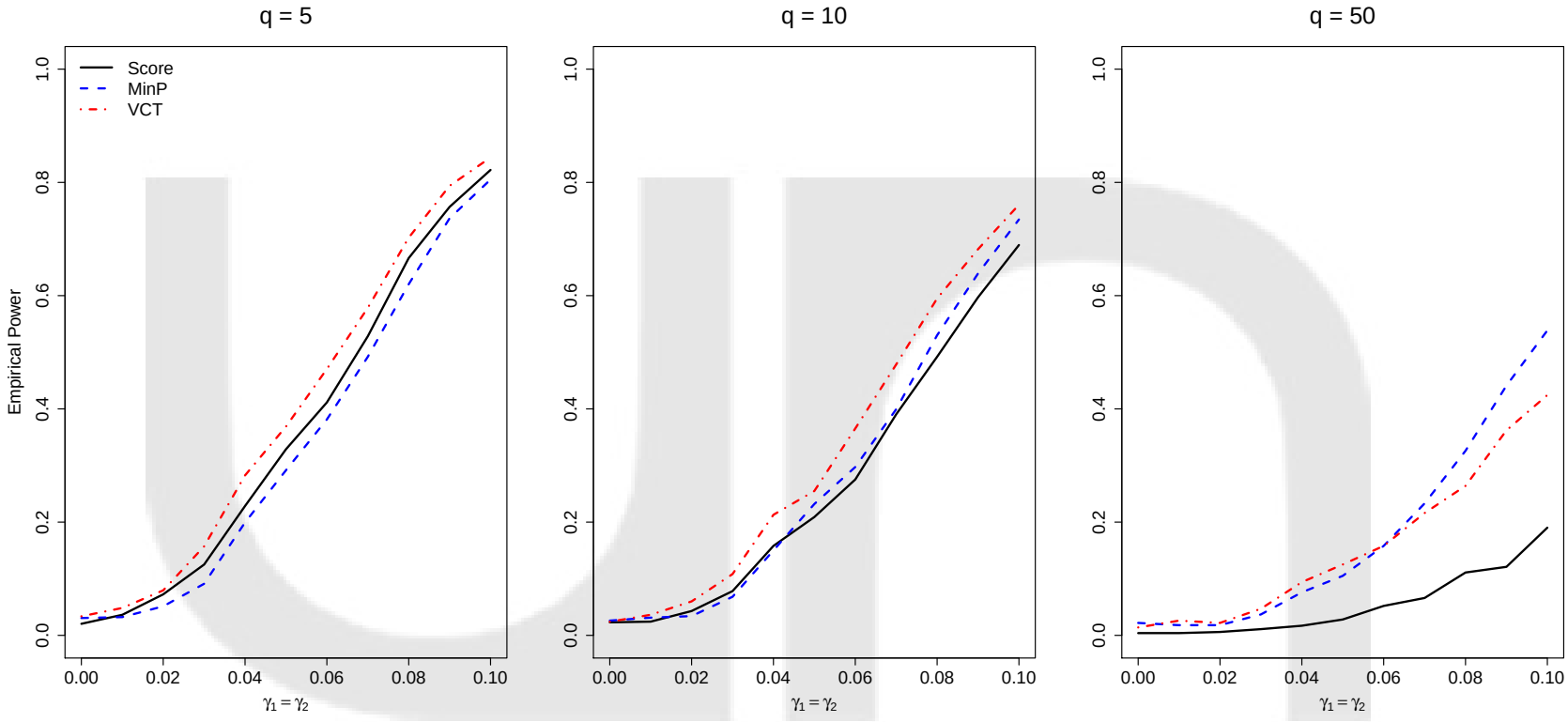
We first compared the empirical type I error of the different methods at 0.05 α-level. To evaluate type I error, we set $\gamma_1 = \gamma_2 = 0$ and varied the number of SNPs q in the gene. The SNPs were either independent or in LD.

Type I error

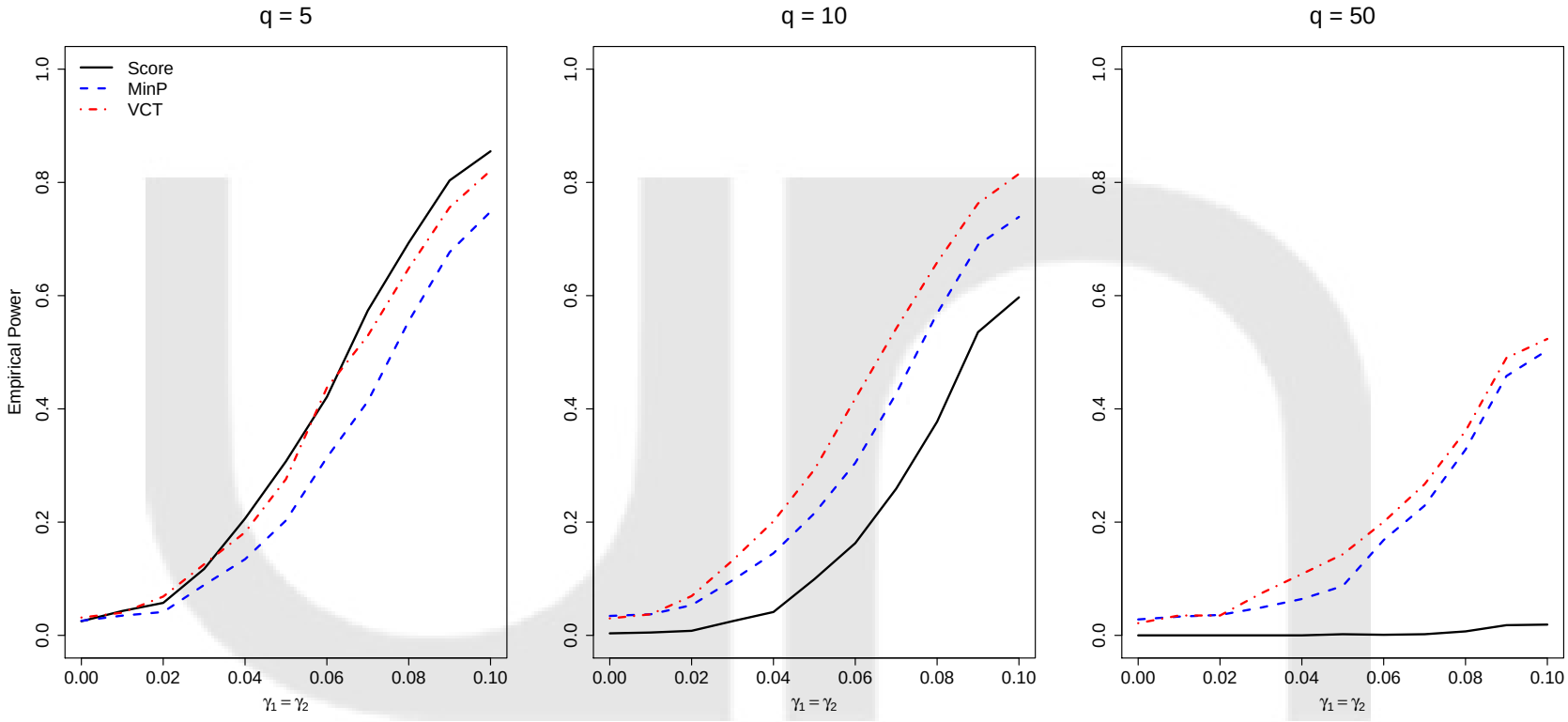
We first compared the empirical type I error of the different methods at 0.05 α-level. To evaluate type I error, we set $\gamma_1 = \gamma_2 = 0$ and varied the number of SNPs q in the gene. The SNPs were either independent or in LD.

SNPs category	q	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_\theta$	$\widehat{\lambda} = 1/\widehat{\sigma}^2_\theta$	Score	MinP	VCT
Independent	5	1.247	0.034	29.4	0.020	0.031	0.034
	10	1.240	0.017	58.8	0.023	0.026	0.024
	50	1.222	0.003	333	0.004	0.022	0.014
LD	5	1.243	0.021	47.6	0.025	0.026	0.031
	10	1.239	0.009	111	0.004	0.034	0.030
	50	1.222	0.002	500	0.000	0.028	0.022

Empirical power for independent SNPs



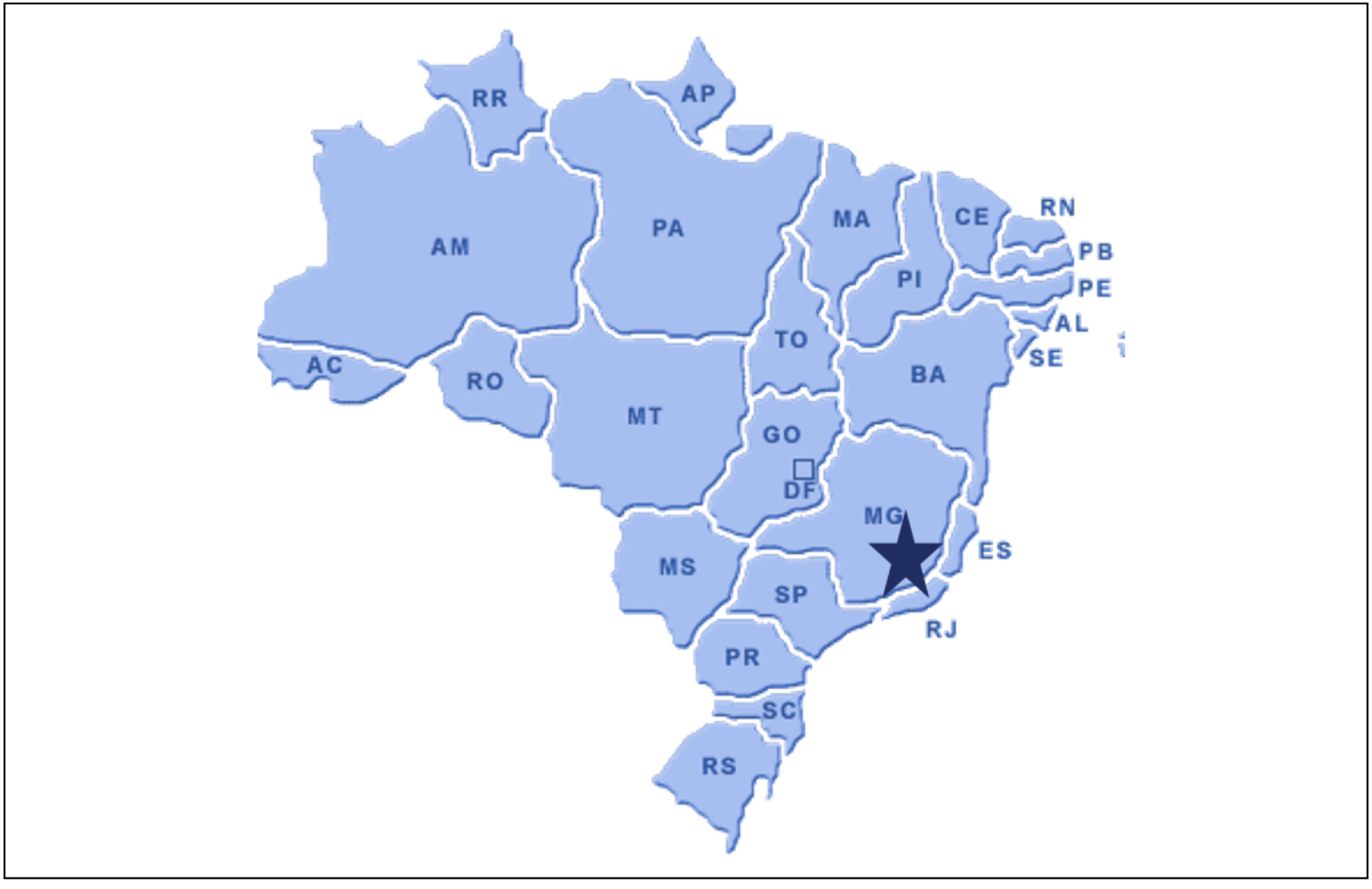
Empirical power for dependent SNPs



Sección 6

Application

Baependi Heart Study



Variables

- *Phenotype:* Type II diabetes
- *Environmental variable:* Body Mass Index (BMI)
- *Genotype:* Were considered the following three genetic regions:
  - Peroxisome-Proliferator-Activated Receptors gamma (PPARG) with 16 variants genotyped;
  - Fat Mass and Obesity associated protein (FTO) with 149 variants genotyped;
  - Cyclin-dependent kinase 5 regulatory subunit associated protein 1-like 1 (CDKAL1) with 186 variants genotyped.
- *Covariates:* Age, sex, and the first two principal components of the entire genotype data of Baependi.

Summary of cases per subjects and families

Gene	Subjects			Families		
	Control	Cases	Total	Control	Cases	Total
PPARG	845	83	928	43	42	85
FTO	712	71	783	47	38	85
CDKAL1	661	69	730	47	38	85

Summary of cases per subjects and families

Gene	Subjects			Families		
	Control	Cases	Total	Control	Cases	Total
PPARG	845	83	928	43	42	85
FTO	712	71	783	47	38	85
CDKAL1	661	69	730	47	38	85

R version 3.3.1 and a processor Intel(R) Core(TM) i5-6500 CPU @ 3.20GHz with a RAM memory 8.00 GB and operating system 64-bits.

Sample size, GLMM parameters, p-values and execution times

Gene	SNPs	Total Subjects	GLMM Parameters			Test	p-Value	α Level	Time (s)
			$\widehat{\sigma}^2$	$\widehat{\sigma}^2_\theta$	$\widehat{\lambda} = 1/\widehat{\sigma}^2_G$				
PPARG	16	928	0.4463	0.0029	344.8276	VCT	0.028	0.05	18.420
						MinP	0.019 *	0.005	100.261
						Score	0.595	0.05	9.025
FTO	149	783	0.3710	0.0033	303.0303	VCT	0.451	0.05	12.958
						MinP	0.031 *	0.0005	2675.907
						Score	0.992	0.05	6.197
CDKAL1	186	730	0.0918	0.0111	90.0901	VCT	0.635	0.05	9.907
						MinP	0.040 *	0.0005	1755.881
						Score	0.915	0.05	4.257

* Compare MinP test p-value with the corresponding corrected α, obtained by dividing 0.05 for the number of effective SNPs (which is equivalent to the number of principal components that reach 99.5% of the their total variation): 10 for *PPARG*, 93 for *FTO* and 92 for *CDKAL1*.

Sección 7

References

References

- Breslow, N. and Clayton, D, 1993. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, **88**, 9-25.
- Coombes, B., Basu, S. and Mcgue, M. , 2017. A combination test for detection of gene-environment interaction in cohort studies. *Genet. Epidemiol.*, **41**, 396-412.
- Chen, H. and Conomos, M., 2016. GMMAT: Generalized Linear Mixed Model Association Tests; R Package Version 0.7. Available online: *https://content.sph.harvard.edu/xlin/dat/GMMAT_user_manual_v0.7.pdf* (accessed on 16 January 2017).
- Lin, X., Lee, S., Chistiani, D. and Lin, X., 2013. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, **14**, 667-681.

References

- Lin, X., Lee, S., Wu, M., Wang, C., Chen, H., Li, Z. and Lin, X., 2016. Test for rare variants by environment interactions in sequencing association studies. *Biometrics*, **72**, 156-164.
- Lin, X, 1997. Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309-326.
- Oliveira, C., Pereira, A., de Andrade, M., Soler, J. and Krieger, J., 2008. Heritability of cardiovascular risk factors in a brazilian population: Baependi heart study. *BMC Med. Genet.*, **9**, 32.
- Shen, X., Alam, M., Fikse, F. and Ronnegard, L., 2013. A novel generalized ridge regression method for quantitative genetics. *Genetics*, **193**, 1255-1268.
- Zhang, D. and Lin, X, 2003. Hyphotesis testing in semiparametric additive mixed models. *Biostatistics* 2003, **4**, 7-74.

References

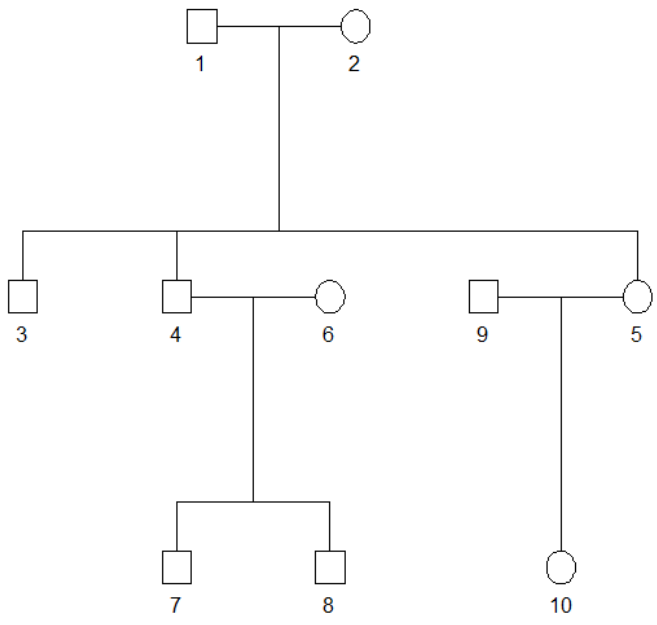
- Satterthwaite, F., 1941. Synthesis of variance. *Psychometrika*, **6**, 309-316.
- Fu, W.J., 2005.Nonlinear GCV and quasi-GCV for shrinkage models. *Journal of Statistical Planning and Inference*, **131**, 333-347.
- Mazo, M. A., Coombes, B. and de Andrade, M., 2017. An Efficient Test for Gene-Environment Interaction in Generalized Linear Mixed Models with Family Data. *International Journal Of Environmental Research And Public Health*, **14**, 1134-1146.

References

- Chen, H., Wang, C., Conomos, M., Stilp, A., Li, Z., Sofer, T., Szpiro, A., Chen, W., Brehm, J., Celedon, J., Redline, S., Papanicolaou, G., Thornton, T., Laurie, C., Rice, K. and Lin, X., 2016. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American journal of human genetics*, **98**, 653-666.
- Leal, S., Yan, K. and Muller-Myhsokb, B, 2005. SimPed: A Simulation Program to Generate Haplotype and Genotype 308 Data for Pedigree Structures. *Hum Hered.*, 119-122.

Family data and kinship matrix

PEDIGREE



KINSHIP MATRIX

Subject	1	2	3	4	5	6	7	8	9	10
1	1.00	0.00	0.50	0.50	0.50	0.0	0.250	0.250	0.0	0.250
2	0.00	1.00	0.50	0.50	0.50	0.0	0.250	0.250	0.0	0.250
3	0.50	0.50	1.00	0.50	0.50	0.0	0.250	0.250	0.0	0.250
4	0.50	0.50	0.50	1.00	0.50	0.0	0.500	0.500	0.0	0.250
5	0.50	0.50	0.50	0.50	1.00	0.0	0.250	0.250	0.0	0.500
6	0.00	0.00	0.00	0.00	0.00	1.0	0.500	0.500	0.0	0.000
7	0.25	0.25	0.25	0.50	0.25	0.5	1.000	0.500	0.0	0.125
8	0.25	0.25	0.25	0.50	0.25	0.5	0.500	1.000	0.0	0.125
9	0.00	0.00	0.00	0.00	0.00	0.0	0.000	0.000	1.0	0.500
10	0.25	0.25	0.25	0.25	0.50	0.0	0.125	0.125	0.5	1.000

Return

Ridge regression estimation

$$g(\boldsymbol{\mu}^{d_2}) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{d}_2 \quad (\boldsymbol{d}_2 = \boldsymbol{K}\boldsymbol{b})$$

Ridge regression estimation

$$g(\boldsymbol{\mu}^{d_2}) = \tilde{\boldsymbol{X}}\beta + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{d}_2 \quad (\boldsymbol{d}_2 = \boldsymbol{K}\boldsymbol{b})$$

$$\begin{aligned} ql(\beta, \boldsymbol{\theta}, \phi_R, \sigma_R) &= -\frac{1}{2}\log\left|\frac{\sigma_R^2}{\phi_R}(\boldsymbol{K}\boldsymbol{K}^T)\boldsymbol{W}_R + \boldsymbol{I}_n\right| \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{n_i} ql_{ij}(\beta, \boldsymbol{\theta}, \phi_R; \tilde{\boldsymbol{d}}_2) - \frac{1}{2}\tilde{\boldsymbol{d}}_2^T\left(\sigma_R^2\boldsymbol{K}\boldsymbol{K}^T\right)^{-1}\tilde{\boldsymbol{d}}_2, \end{aligned}$$

where $\tilde{\boldsymbol{d}}_2$ is choosen to maximize the sum of the last two terms, $\boldsymbol{W}_R = \text{diag}\left\{\omega_{ij}/\left[\nu(\mu_{ij}^{d_2})g'(\mu_{ij}^{d_2})^2\right]\right\}$ and

$$ql_{ij}(\beta, \boldsymbol{\theta}, \phi_R; \boldsymbol{d}_2) = \int_{Y_{ij}}^{\mu_{ij}^{d_2}} \frac{\omega_{ij}(Y_{ij} - \mu)}{\phi_R\nu(\mu)}d\mu.$$

Ridge regression estimation

Ridge regression estimator of $\boldsymbol{\theta}$, is obtained by minimazing the function

$$[ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R) \times \phi_R] - \frac{1}{2}\lambda \boldsymbol{\theta}^T \boldsymbol{\theta}.$$

where λ is a penalizing factor.

$$\begin{bmatrix} \tilde{\boldsymbol{X}}^T \boldsymbol{W}_R \tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^T \boldsymbol{W}_R & \tilde{\boldsymbol{X}}^T \boldsymbol{W}_R \\ \boldsymbol{W}_R \tilde{\boldsymbol{X}} & \boldsymbol{W}_R + \lambda(\boldsymbol{G}\boldsymbol{G}^T)^{-1} & \boldsymbol{W}_R \\ \boldsymbol{W}_R \tilde{\boldsymbol{X}} & \boldsymbol{W}_R & \boldsymbol{W}_R + \frac{\phi_R}{\sigma_R^2}(\boldsymbol{K}\boldsymbol{K}^T)^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ \boldsymbol{d}_1 \\ \boldsymbol{d}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}^T \boldsymbol{W}_R \tilde{\boldsymbol{Y}} \\ \boldsymbol{W}_R \tilde{\boldsymbol{Y}} \\ \boldsymbol{W}_R \tilde{\boldsymbol{Y}} \end{pmatrix}$$

where $\boldsymbol{d}_1 = \boldsymbol{G}\boldsymbol{\theta}$. Return